MODEL REGRESI NONPARAMETRIK DENGAN PENDEKATAN SPLINE TRUNCATED

Rahmat Hidayat[1], Yuliani[2], Marwan Sam[3]
Institut Teknologi Sepuluh Nopember[1], Universitas Cokroaminoto Palopo[2,3]

dayatmath@gmail.com[1]

Analisis regresi merupakan salah satu analisis di dalam statistika yang digunakan untuk mengestimasi pola hubungan antara variabel prediktor (x) dengan variabel respon (y). Diberikan data berpasangan (x_j, y_j) dengan hubungan antara x_j dengan y_j diasumsikan mengikuti model regresi nonparametrik $y_j = f(x_j) + \varepsilon_j$, $j = 1, 2, ..., n$. Kurva regresi dihampiri dengan fungsi spline *g* dengan *r* buah titik knot *K*. Estimasi untuk kurva regresi spline diberikan oleh $\hat{f}(x_j) = A[K]\underset{\sim}{y}$. Estimator spline sangat bergantung pada titik knot dan sifat estimator spline adalah linier.

Kata kunci: nonparametrik, *spline*

1. Pendahuluan

Analisis regresi dipakai berkenaan dengan studi pola hubungan antar satu atau lebih variabel prediktor dengan maksud menaksir atau memprediksi variabel respon. Selain untuk mengetahui pola hubungan, analisis regresi juga bisa dipakai untuk memprediksi. Analisis regresi dalam mengestimasi kurva regresi terdapat tiga pendekatan, yaitu pendekatan regresi parametrik, regresi nonparametrik dan regresi semiparametrik. Dalam pendekatan regresi parametrik terdapat asumsi yang sangat kuat dan kaku yaitu bentuk kurva regresi diketahui misalnya linier, kuadratik, kubik, polinomial derajat *p*, eksponen dan lain-lain. Disamping itu, diperlukan pengetahuan masa lalu tentang karakteristik data agar memperoleh pemodelan yang baik. Tujuan utama dalam analisis regresi adalah mencari bentuk estimasi kurva regresi. Dalam model regresi parametrik estimasi kurva regresi ekuivalen dengan estimasi terhadap parameter-parameter dalam model [1].

Pendekatan model regresi parametrik memiliki sifat yang sangat baik dari pandangan Statistika Inferensia seperti sederhana, mudah interpretasinya, parsimoni, tak bias, estimator linier, efisien, konsisten, dan *Best Linier Unbiased Estimator* (BLUE). Walaupun sangat handal dari sisi inferensia, akan tetapi pendekatan ini sangat membutuhkan terpenuhi asumsiasumsi yang disyaratkan. Tidak semua permasalahan pola hubungan dapat didekati dengan regresi parametrik karena tidak semua permasalahan mempunyai informasi bentuk hubungan atau kurva regresi antar variabel respon dengan variabel prediktor. Jika dipaksakan dengan pendekatan regresi

parametrik maka akan memberikan kesimpulan yang menyesatkan. Oleh karena itu, ada pilihan dengan mengggunakan pendekatan nonparametrik yang tidak memerlukan informasi pola hubungan antar variabel.

Dalam model regresi nonparametrik bentuk kurva regresi diasumsikan tidak diketahui. Kurva regresi hanya diasumsikan halus (*smooth*) dalam arti termuat di dalam suatu ruang fungsi tertentu (ruang *Hilbert*, ruang *Sobolev*, ruang *HilbertSobolev*, ruang *Banach*, ruang fungsi *kontinu*, ruang *Entropi*, dan lain-lain) [1]. Perbedaan antara parametrik dengan nonparametrik adalah dalam pendekatan parametrik data cenderung dipaksa untuk mengikuti pola tertentu, sedangkan pendekatan nonparametrik data diberi keleluasaan untuk mencari pola kurva regresinya sendiri sehingga sangat fleksibel dan obyektif. Beberapa model regresi nonparametrik yang banyak digunakan antara lain Histogram, Kernel, Spline, Polinominal Lokal, Deret Ortogonal, Deret Fourier, kNN, Neural Network (NN), Wavelets, MARS, dan yang lain. Semua model memiliki kelebihan dan kekurangan serta motivasi sendiri dalam memodelkan pola [1]. Dari beberapa model regresi nonparametrik tersebut yang paling popular adalah Spline.

Penelitian secara teoritis juga banyak berkembang antara lain dilakukan [2]; [3]; [4]. Spline merupakan potongan-potongan polinomial yang memiliki sifat tersegmen dan kontinu (*truncated*). Regresi spline *polynomial truncated* digunakan karena mempunyai kelebihan yaitu model ini cenderung mencari sendiri estimasi data kemanapun pola data tersebut bergerak. Kelebihan ini terjadi karena dalam spline terdapat titik-titik knot, yaitu titik perpaduan bersama yang menunjukkan terjadinya perubahan pola perilaku data [5]. Dengan titik knot ini, spline dapat memberikan fleksibilitas yang lebih baik dari pada polinomial, sehingga memungkinkan untuk menyesuaikan diri secara efektif terhadap karakteristik lokal. Alasan lain memilih regresi spline *polynomial truncated* karena objektif dan optimasinya menggunakan metode *least square* sehingga secara matematik mudah, sederhana, dan baik dalam membantu inferensi statistik. [6] dalam artikelnya menyarankan menggunakan regresi spline polinomial jika plot data tidak jelas, standar deviasi besar, dan untuk lebih mudah.

[7] memperlihatkan dalam regresi nonparametrik spline data *cross section* dan satu respon, bahwa jika nilai parameter penghalus sangat kecil maka akan memberikan estimator kurva regresi yang sangat kasar. Sebaliknya, jika nilai parameter penghalus sangat besar maka akan dihasilkan estimator kurva regresi

nonparametrik yang sangat mulus. Akibatnya dalam estimator spline untuk data *cross section* perlu dipilih parameter penghalus yang optimal agar diperoleh estimator yang paling sesuai untuk data. Dalam penelitian ini, akan dilakukan estimasi kurva regresi nonparametrik menggunakan pendekatan spline truncated serta bagaimana sifat-sifat dari estimator yang diperoleh.

Analisis Regresi

Analisis regresi merupakan suatu analisis statistika yang digunakan untuk mengetahui persamaan pola hubungan antara variabel prediktor dan variabel respon. Terdapat dua pendekatan estimasi model dalam analisis regresi, yaitu regresi parametrik dan regresi nonparametrik. Pendekatan regresi parametrik digunakan jika bentuk kurva regresi diketahui. Jika pola hubungan data membentuk pola linear maka digunakan pendekatan regresi parametrik linear. Jika pola hubungan data membentuk pola kuadrat maka digunakan pendekatan regresi kuadratik, dan lain-lain [8]. Bentuk pola hubungan dapat diidentifikasi berdasarkan pada informasi masa lalu atau *scatter plot* data [9].

Analisis Regresi Linier

Regresi linier merupakan metode statistika yang digunakan untuk memodelkan hubungan antara variabel respon dengan satu atau lebih variabel prediktor. Apabila banyaknya variabel prediktor hanya ada satu, disebut sebagai regresi linier sederhana, sedangkan apabila terdapat lebih dari satu variabel prediktor, disebut sebagai regresi linier berganda. Analisis regresi mampu mendeskripsikan fenomena data melalui terbentuknya suatu model hubungan yang bersifatnya numerik. Jika *x* adalah variabel prediktor dan *y* adalah variabel respon, maka terdapat hubungan fungsional antara *x* dan *y*. Jika dibuat secara matematis hubungan itu dapat dijabarkan sebagai berikut.

$$y = f\left(x\right) + \varepsilon \tag{0.1}$$

dimana y adalah adalah variabel respon yang merupakan suatu fungsi dari variabel prediktor x, serta ε adalah *error* acak yang diasumsikan mengikuti distribusi normal [10].

Spline dalam Regresi Nonparametrik

Spline dalam regresi nonparametrik mempunyai sifat fleksibilitas yang tinggi dan mempunyai kemampuan mengestimasi perilaku data yang cenderung berbeda pada interval yang berlainan ([8];[11]). Kemampuan mengestimasi perilaku data ini ditunjukkan oleh fungsi *truncated* (potongan-potongan) yang melekat pada estimator

dan potongan-potongan tersebut yang disebut titik knot. Titik knot merupakan titik perpaduan bersama yang menunjukkan perubahan pola perilaku fungsi pada selang yang berbeda. Spline merupakan salah satu jenis *piecewise* polinomial, yaitu polinomial yang memiliki sifat tersegmen. Sifat tersegmen ini memberikan fleksibilitas lebih dari polinomial biasa, sehingga memungkinkan untuk menyesuaikan diri secara lebih efektif terhadap karakteristik lokal suatu fungsi atau data. Dalam fungsi spline terdapat titik knot yang merupakan titik perpaduan yang menunjukkan perubahan perilaku kurva pada selang yang berbeda [9]. Fungsi spline berderajat *m* adalah sebarang fungsi yang secara umum dapat disajikan dalam bentuk sebagai berikut:

$$f\left(x_i\right)=\sum_{j=0}^{m}\beta_j x_i^j+\sum_{j=1}^{J}\beta_{j+m}\left(x_i-k_j\right)_+^m \qquad (0.2)$$

dengan β_j adalah konstanta riil, dan

$$\left(x_i-k_j\right)_+^m=\begin{cases}\left(x_i-k_j\right)^m; x\geq k_j\\ 0 \quad ; x<k_j\end{cases}$$

Jika *m* = 1,2, dan 3 diperoleh berturut-turut spline linear, spline kuadratik dan spline kubik serta k_j adalah titik knot.

Apabila diasumsikan *error* ε_i berdistribusi normal independen dengan ratarata nol dan variansi σ^2, maka y_i pada model regresi juga berdistribusi normal dengan rata-rata $f\left(x_i\right)$ dan variansi σ^2. Akibatnya diperoleh estimasi untuk parameter β dengan menggunakan metode *least square,* yaitu dengan meminimumkan jumlah kuadrat *error*-nya adalah sebagai berikut,

$$\sum_{i=1}^{n}\varepsilon_i^2=\left(y_i-f\left(x_i\right)\right)^2$$

$$=\left(y_i-\left(\sum_{j=0}^{m}\beta_j x_i^j+\sum_{j=1}^{J}\beta_{j+m}\left(x_i-k_j\right)_+^m\right)\right)^2$$

Dengan penyajian matriks, diperoleh:

$$\sum_{i=1}^{n}\varepsilon_i^2=\underset{\sim}{\varepsilon}'\underset{\sim}{\varepsilon}$$

$$=\left(\underset{\sim}{y}-X\underset{\sim}{\beta}\right)'\left(\underset{\sim}{y}-X\underset{\sim}{\beta}\right)$$

$$=\underset{\sim}{y}'\underset{\sim}{y}-2\underset{\sim}{\beta}'X'\underset{\sim}{y}+\underset{\sim}{\beta}'X'X\underset{\sim}{\beta}$$

Bila persamaan di ats diturunkan terhadap vektor $\underset{\sim}{\beta}'$ dan hasilnya disamakan dengan nol maka didapat:

$$\hat{\underset{\sim}{\beta}} = \left(X'X\right)^{-1} X'\underset{\sim}{y} \tag{0.3}$$

dengan:

$$X = \begin{bmatrix} 1 & x_1 & \cdots & x_1^m & \left(x_1 - k_1\right)_+^m & \cdots & \left(x_1 - k_j\right)_+^m \\ 1 & x_2 & \cdots & x_2^m & \left(x_2 - k_1\right)_+^m & \cdots & \left(x_2 - k_j\right)_+^m \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_n & \cdots & x_n^m & \left(x_n - k_1\right)_+^m & \cdots & \left(x_n - k_j\right)_+^m \end{bmatrix}$$

Estimator Spline dalam Regresi Nonparametrik

Fungsi spline merupakan jumlahan dari fungsi polinomial dengan suatu fungsi *truncated.* Dalam bagian ini dibahas tentang model regresi nonparametrik, dimana estimasi kurva f dilakukan dengan menggunakan spline. Diberikan data berpasangan $\left(x_j, y_j\right)$ dan hubungan antara x_j dengan y_j diasumsikan mengikuti model regresi nonparametrik:

$$y_j = f\left(x_j\right) + \varepsilon_j,\ j = 1, 2, ..., n$$

Di dalam penelitian ini, dilakukan suatu kajian dengan kurva regresi f dihampiri dengan fungsi spline f dengan knot *K.* Dalam bentuk matrik disajikan sebagai berikut:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} f\left(x_1\right) \\ f\left(x_2\right) \\ \vdots \\ f\left(x_n\right) \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Apabila model regresi spline disajikan dalam bentuk matriks, diperoleh:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^m & \left(x_1 - K_1\right)_+^m & \cdots & \left(x_1 - K_r\right)_+^m \\ 1 & x_2 & x_2^2 & \cdots & x_2^m & \left(x_2 - K_1\right)_+^m & \cdots & \left(x_2 - K_r\right)_+^m \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^m & \left(x_n - K_1\right)_+^m & \cdots & \left(x_2 - K_r\right)_+^m \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_m \\ \beta_1 \\ \vdots \\ \beta_r \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

atau dapat ditulis dengan

$$\underset{\sim}{y} = X\left[K_1, K_2, ..., K_r\right]\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$$

Selanjutnya, estimasi parameter $\underset{\sim}{\beta} = \begin{pmatrix} \alpha_0 & \alpha_1 & \alpha_2 & \cdots & \alpha_m & \beta_1 & \cdots & \beta_r \end{pmatrix}'$ diperoleh dengan metode *least square*, dengan menyelesaikan optimasi:

$$\underset{\underset{\sim}{\beta}\in R^{m+1+r}}{Min}\left\{\underset{\sim}{\varepsilon}'\underset{\sim}{\varepsilon}\right\} = \underset{\underset{\sim}{\beta}\in R^{m+1+r}}{Min}\left\{\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\\ \vdots\\ \varepsilon_n\end{pmatrix}'\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\\ \vdots\\ \varepsilon_n\end{pmatrix}\right\}$$

$$= \underset{\underset{\sim}{\beta}\in R^{m+1+r}}{Min}\left\{\left(\underset{\sim}{y} - X\left[K_1, K_2, ..., K_r\right]\underset{\sim}{\beta}\right)'\left(\underset{\sim}{y} - X\left[K_1, K_2, ..., K_r\right]\underset{\sim}{\beta}\right)\right\}$$

Jumlah kuadrat *error* dengan penjabaran matriksnya diberikan sebagai berikut:

$$\sum_{i=1}^{n}\varepsilon_i^2 = \underset{\sim}{\varepsilon}'\underset{\sim}{\varepsilon}$$

$$= \left(\underset{\sim}{y} - X\left[K_1, K_2, ..., K_r\right]\underset{\sim}{\beta}\right)'\left(\underset{\sim}{y} - X\left[K_1, K_2, ..., K_r\right]\underset{\sim}{\beta}\right)$$

$$= \left(\underset{\sim}{y} - X\left[K\right]\underset{\sim}{\beta}\right)'\left(\underset{\sim}{y} - X\left[K\right]\underset{\sim}{\beta}\right)$$

$$= \underset{\sim}{y}'\underset{\sim}{y} - 2\underset{\sim}{\beta}'X\left[K\right]'\underset{\sim}{y} + \underset{\sim}{\beta}'X\left[K\right]'X\left[K\right]\underset{\sim}{\beta}$$

Bila persamaan di atas diturunkan terhadap vektor β' dan hasilnya disamakan dengan nol, diperoleh:

$$\frac{\partial\left(\underset{\sim}{\varepsilon}'\underset{\sim}{\varepsilon}\right)}{\partial\underset{\sim}{\beta}'} = \frac{\partial\left(\underset{\sim}{y}'\underset{\sim}{y} - 2\underset{\sim}{\beta}'X\left[K\right]'\underset{\sim}{y} + \underset{\sim}{\beta}'X\left[K\right]'X\left[K\right]\underset{\sim}{\beta}\right)}{\partial\underset{\sim}{\beta}'} = 0$$

$$\underset{\sim}{\hat{\beta}} = \left(X\left[K\right]'X\left[K\right]\right)^{-1}X\left[K\right]'\underset{\sim}{y}$$

dengan,

$$X\left[K\right] = X\left[K_1, K_2, ..., K_r\right]$$

$$= \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^m & \left(x_1 - K_1\right)_+^m & \cdots & \left(x_1 - K_r\right)_+^m \\ 1 & x_2 & x_2^2 & \cdots & x_2^m & \left(x_2 - K_1\right)_+^m & \cdots & \left(x_2 - K_r\right)_+^m \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^m & \left(x_n - K_1\right)_+^m & \cdots & \left(x_2 - K_r\right)_+^m \end{pmatrix}$$

Akibatnya, estimasi untuk kurva regresi spline dengan knot *K* diberikan oleh :

$$\hat{f}\left(x_i\right) = X\left[K\right]\hat{\underset{\sim}{\beta}}$$
$$= X\left[K\right]\left(X\left[K\right]' X\left[K\right]\right)^{-1} X\left[K\right]' \underset{\sim}{y}$$

Sehingga diperoleh:

$$\hat{f}\left(x_i\right) = A\left[K\right]\underset{\sim}{y}$$

dengan $A\left[K\right] = X\left[K\right]\left(X\left[K\right]' X\left[K\right]\right)^{-1} X\left[K\right]'$ merupakan fungsi dari titik knot dan $K = \left(K_1, K_2, ..., K_r\right)'$ merupakan titik-titik knot.

Sifat Kelinieran Estimator

Model regresi nonparametrik spline dapat ditulis dalam bentuk matriks berikut.

$$\underset{\sim}{y} = X\left[K\right]\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$$

dengan $K = \left(K_1, K_2, ..., K_r\right)$. Jika dituliskan $\underset{\sim}{f} = X\left[K\right]\underset{\sim}{\beta}$, dengan $X\left[K\right]$ merupakan matriks fungsi dari *K*, maka diperoleh:

$$\underset{\sim}{y} = \underset{\sim}{f} + \underset{\sim}{\varepsilon}$$

dan

$$\hat{\underset{\sim}{f}} = X\left[K\right]\hat{\underset{\sim}{\beta}}$$
$$= X\left[K\right]\left(X\left[K\right]' X\left[K\right]\right)^{-1} X\left[K\right]' \underset{\sim}{y}$$
$$= A\left[K\right]\underset{\sim}{y}$$

Berdasarkan persamaan di atas, terlihat bahwa estimator spline $\hat{\underset{\sim}{f}}$ merupakan estimator yang linier. Kelinieran ini dapat memberikan kemudahan bagi peneliti dalam membentuk statistik inferensi untuk pendekatan spline.

Kesimpulan

1. Jika diketahui model regresi nonparametrik:

   $$y_j = f\left(x_j\right) + \varepsilon_j,\ j = 1, 2, ..., n$$

   Kurva regresi f dihampiri dengan fungsi spline dengan r buah titik knot K, diperoleh

   $$\hat{\underset{\sim}{f}} = A\left[K_1, K_2, ..., K_r\right]\underset{\sim}{y} = A\left[K\right]\underset{\sim}{y}$$

dengan $A[K] = X[K]\left(X[K]' X[K] \right)^{-1} X[K]'$ merupakan fungsi dari titik-titik knot $K = (K_1, K_2, ..., K_r)'$.

2. Sifat estimator yang diperoleh adalah linier.

Daftar Pustaka

[1] Budiantara, I.N., Lestari, B., Islamiyati, A., Pemilihan Knot Optimal dalam Estimator Spline Terbobot pada Regresi Nonparametrik Heteroskedastik Data Longitudinal, *Seminar Nasional Statitsika IX, Institut Teknologi Sepuluh Nopember, Surabaya.* 2009.

[2] Doksum,K., dan Koo,Y.J., "On Spline Estimators and Prediction Intervals in Nonparametric Regression", *Computational Statistics and Data Analysis,* 2000, vol. 35, 67 – 82, 2000.

[3] Wand, M.P., *"A Comparison of Regression Spline Smoothing Procedures",* Departments of Biostatistics, School of Public Health, Harvard, 2005.

[4] Huang, Z.J., "Local Asymptotic for Polynomial Spline Regression", *The Annual Statistics,* vol. 31, no. 5, 1600 – 1635, 2003.

[5] Eubank, R.L., *Spline Smoothing and Nonparametric Regression 2nd Edition,* Marcel Dekker, New York, 1999.

[6] Hurley, D., Hussey,J., McKeown,R., dan Addy,C., "An Evaluation of Splines in Linier Regression", South Carolina Central Cancer Registry, Columbia, 2005.

[7] Wahba, G., "Spline Models for Observational Data", *SIAM, CBMS-NSF Regional Conference Series and Applied Mathematics, Philadelphia.* 1990.

[8] Eubank, R.L., "*Spline Smoothing and Nonparametric Regression"*, Mercel Dekker, New York, 1988.

[9] Hardle, W., "*Applied Nonparametric Regression"*, Cambridge University Press, New York, 1990.

[10] Draper, N.R dan Smith. H., "*Applied Regression Analysis (third edition)*", Canada: John and Wileys & Sons,inc, 1998.

[11] Budiantara, I.N., "Model Spline dengan Knots Optimal", *Jurnal Ilmu Dasar, FMIPA Universitas Jember, 7, 77-85*, 2006.